VSMS- A Survey on Voice based SMS system

Kamakshi Thakkar[#1] , Isha Havaldar [#2] , Darshan Gondalia [#3], Aarti Sahitya[#4]
K. J. Somaiya Institute of Engineering and Information Technology, Sion. Mumbai-22.
Department of Computer Science
kamakshi.thakkar@somaiya.edu, isha.havaldar@somaiya.edu, darshan.gondalia@somaiya.edu,
aarti.sahitya@somaiya.edu

Abstract

Speech recognition has been a wide area of research and development over years.Due to easy usability and high efficiency , it has gained a great importance in the field of technology. Before a decade speech recognition was difficult but with the growth of technology many new algorithms , techniques and tools have been implemented successfully, some of which include HMM framework , deep learning algorithms, etc. SMS has been the second largest element used for communication. This paper elaborates an idea of building an 'SMS System' for senior citizens, illiterate individuals,etc. in order to make the process of communication easier and better for them. Right from selection of the contacts to sending the message, the whole process will be performed with the help of voice commands. The basic objective is to develop an offline SMS system in which the speech recognition would be performed in an efficient manner. Multiple contact selection, low frequency voice recognition, external noise elimination would be some features of the system.

Introduction

With advances in the field of information technologies, mobile devices have become important tools to connect individuals over long distances within a few seconds. People with physical disabilities or senior citizens find these technologies as a curse because most of them cannot use it or even if they can, they are unaware of their functionalities. Such devices and its services are often unavailable because they require proper adaptive tools and some special interfaces in order to use mobile phone devices in an conventional manner. There are times when physically handicapped want to communicate but are unable to do so through any of its medium that is neither by call nor by SMS. The Voice based SMS system will provide a platform for such users to send messages through simple voice commands which will reduce the efforts of the handicapped as well as senior citizens and thus would be of great use to them.

While there are various algorithms for speech to text conversion available, each of them do not have proper word conversion accuracy. Also, the noise elimination is improper and there is a narrow range of voice frequency detection i.e it is unsuitable for constantly changing modulation of sound. There are cases where the speakers may have different accents, the pronunciation of a particular word may vary from person to person or there may be difference in the style and rates in which a particular person speaks. In such cases, filtering noise and bringing the frequency of sound to a particular level becomes important as the conversion of speech to text depends on these early stage filtrations and modifications. In addition to human errors and variations caused by human voice, factors like the background noise, echos, the different types of microphones used and the recording devices cause problems in the conversion of speech to text.

Literature Survey:

Various papers on speech to text conversions were studied and certain observations were made about how the existing system is and how it works.

The first paper highlights the use of Kalman Filter and also states the use of it. The main purpose of the Kalman Filter is elimination of background noise to enhance the quality of the words spoken. The main objective of this paper was to invent a new system that would recognize speech in better

way than the HMM model which is used at a large scale in various applications.

The second paper states the objective of speech recognition system which would extract the voice, characterize it and then recognize the information about speech. It uses MFCC i.e Mel-Frequency Cepstral Coefficient feature extraction technique. Initially Voice signal is compressed into features and then the features are used for recognition. MFCC is used to filter out the background voice in the input voice command. It takes place in two phases Training and Testing phase.

The third paper concentrates on the division of the speech recognition process in different phases and gives detailed explanation of each phase. The focus of this system is the development of an online speech-to-text engine. It provides a future scope of developing a system which will process words of different mother tongues.

The fourth paper discusses about how deep learning algorithms of machine learning are used instead of the usual Gaussian mixtures. The main objective of this paper is to apply deep learning algorithms which include (DNN) deep neural networks and (DBN) deep belief networks for automatic and continuous speech recognition.

Among the various techniques and models used, Hidden Markov Model is known to be the best. It is known as the base platform of Speech recognition as most of the developers use it as the native model. It is used in the speech recognition phase which is known to be the most important as it performs the conversion of speech to text. The reason why HMM is used in such a wide range is because a speech signal can be divided into short-time stationary signal and it can be approximated while it is divided into intervals. Each divided part of the signal is considered to be a state which consists of many different hidden states. Then each HMM state utilizes a mixture of Gaussian to model a spectral representation of the sound wave. However, one of the main drawbacks of Gaussian mixture models is that they are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. Also HMM uses MFCC in which the frequency bands are equally spaced on the scale. It approximates the response of the human voice more accurately than the linearly spaced frequency bands like in MFC.

Methodology

Workflow of a system is the most important aspect of any system as it provides the roadmap of developing the system. It acts as a blueprint for the developer while implementing the system. The overall design of our system is shown if fig. below .As shown the whole system is divided into three phases. Each phase has its own purpose and describes the work involved in it.The whole system works serially. All the three phases are equally important as each one of them is dependent on the previous one. The output of the previous phase acts as the input to the next phase. Thus, implementation can be considered to be the most critical stage in developing a successful new system and in giving the user confidence that the new system will work and be effective. The diagramatic representation of the three phases are as follows:

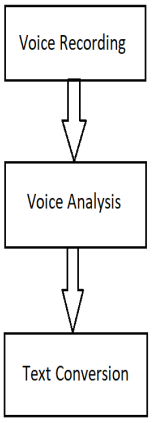

Fig: System Architecture

The above diagram represents the overall architecture of the proposed system . The functionality of each of these three phases are as follows:

1.Voice Recording Phase: In this phase the voice command is recorded using the audio source and stored for future use. A number of classes and methods were used in this phase and involves

processing of the recorded voice and storing it as well. The classes used were the MediaRecorder class and the MediaPlayer class. These methods are used to record an audio and play the recorded audio. The methods used under the MediaRecorder class to record the audio includes setAudioSource(), setOutputFormat(), setAudioEncoder(), setOutputFile(), prepare(), start(), stop(). The methods used under the MediaPlayer class to play the recorded audio includes setDataSource(), prepare(), start(), stop(), release(). Each of these method has its own purpose, setAudioSource() method is used to set the input source from which the signal would be feeded. setOutputFormat() is used to set the audio format in which we want to store the signal. setAudioEncoder() is used to select the encoding algorithm through which we want the input audio must be encoded. setOutputFile() is used to select the file in which the fetched voice command need to be stored. prepare() is used to allocate the resources and keep the system ready for recording the voice command.start() is used to start recording the command and stop() is used to end the recording and deallocate the allocated resources.

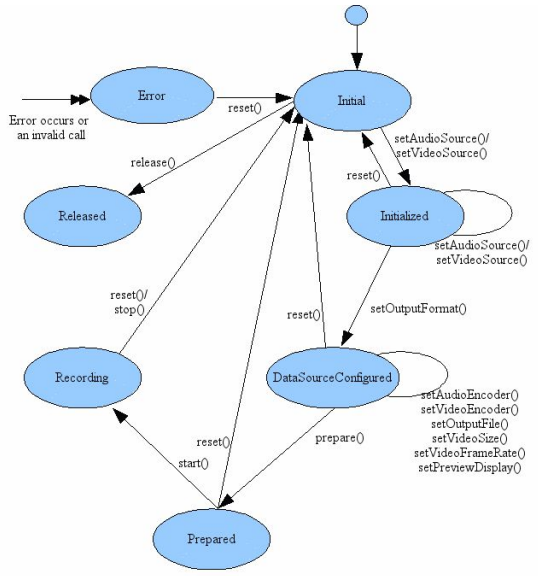

fig: MediaRecorder state diagram

2.Voice Analysis Phase: In this phase the recorded voice is filtered and the unwanted voice is removed. This phase deals with deletion of the noise, echos and also the background gitter of things around. To do this we have used NoiseSuppresor class, AcousticEchoCanceler class and AutomaticGainControl class. Each of these classes share the same methods namely create(int sessionid) and isAvailable(). Both of the methods are static and can be called using the Classname. create method is used to create the instance of the class through which it is called. It intakes a parameter of session id which represents the unique identification number assigned to each of the audio files by the system.Thus the return type of create method is the object of the class calling the create method. isAvailable() is used to check whether the requested class instance is free or not. It returns true if the instance is available or else it would return false. After creating the instance of the required class it is attached to the signal which needs to be filtered using getEnabled() method.It attaches the instance to the voice source and performs the required filtration where and when required. VoiceSupresor class instance remove the unwanted noise and filters the command.AcousticEchoCanceler removes the echo if the command echo gets created during the recording phase. AutomaticGainControl is used to balance the frequency of the voice signal i.e. it amplifies the frequency if it is below the specified band or it would reduce the frequency if it is higher than the specified band.

3.Text Conversion Phase: It would be the final output phase of the system . It would provide us the desired output required by the user. It would convert the filtered voice sample in to text format. HMM methodology would be the bridge which would be used for conversion. It would classify the filtered voice into small samples which would be stationary in nature. These samples are then classified and the hidden samples are identified for recognition. The converted text would be then supplied to the messaging application through intents and on the command the SMS would be sent.

Proposed System:

There are many systems available in the market currently but each of these applications have some or the other drawbacks within them. Our System has some features embedded with the functionality of the system which are unique from the rest of the as-is system available in today's world. The available systems mostly are online in nature i.e. the conversion is performed on some remote location and Internet is mandatorily required for performing conversion.Also these systems do not allow to send SMS to multiple contacts at the same time. These would be the foundation of the developing system. Our system would be fully an offline system which would entirely function on the user side. The overall functionality would be installed while the application gets installed for the first time in the user's device. Preprocessing of the lower frequency voice signal would be performed and then the conversion would take place. The overall user interaction towards the application can be diagrammatically represented by the following fig.

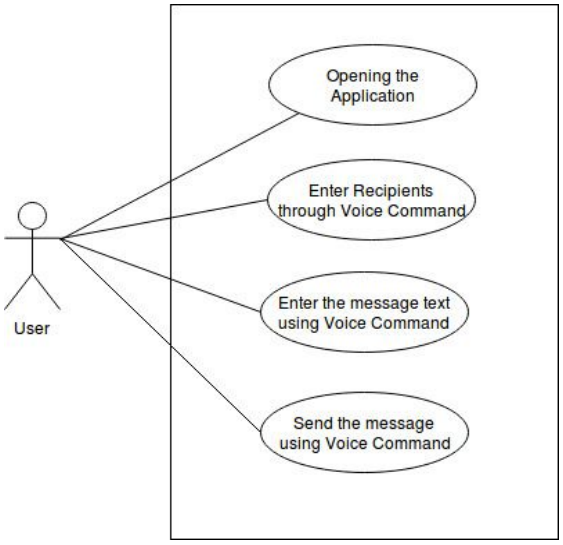

Fig. Use-Case Diagram

Conclusion

In this paper, we analyzed the problems faced by senior citizens and have made the SMS system which will simplify the task of sending messages and also provide a better and easier platform for communication through voice commands. Our system would work in three phases namely the voice recording phase, voice analysis phase and text conversion phase. The voice recording phase uses the MediaRecorder class and MediaPlayer class for recording and storage of the voice commands. In the voice analysis phase, the background noise, echos and the frequency standardization would be done. This would be performed using the NoiseSuppresor class, AcousticEchoCanceler class and AutomaticGainControl class. The last phase involves the conversion of speech to text using the HMM based algorithm. The proposed system will be completely offline i.e it will not require internet connection for its functionality. Also, multiple contact selection and frequency standardization would be some of the additional features of the developing system. Thus we can conclude that our proposed system will be acting as a boon for the senior citizens, illiterates as well as physically handicapped individuals.

Future Scope

Additional features can be added in the current system nearly in future. One of it is that the entire system can be simulated i.e right from starting the application from the background to sending the message will be done through voice commands. Furthermore, multiple languages can be embedded to expand the system to remove language dependency. This system can also be expanded by including the facility of MMS communication.

References

[1]. Neha Sharma, Shipra Sardana, "A real time speech to text conversion system using bidirectional Kalman filter in Matlab," IEEE, 2353-2357, September 2016.
[2]. Yogita Ghadage, Sushama Shelke, "Speech to text conversion for multilingual languages," IEEE, 0236-0240, April 2016.

[3]. Prachi Khilari, Prof. Bhope V. P., "A review on speech to text conversion methods," IJARCET, 3067-3072, July 2015.
[4]. Yan Zhang, Andrew Ng, "Speech recognition using deep learning algorithms", Springer, September 2015.
[5]. V.Naresh, B.Venkataramani, Abhishek Karan and J.Manikandan, "PSOC based isolated speech recognition system," IEEE Internationa Conference on Communication and Signal Processing, pp 693-697,April 3-5, 2013, India.
[6]. Taabish Gulzar, Anand Singh, Dinesh Kumar Rajoriya and Najma Farooq, "A Systematic Analysis of Automatic Speech Recognition: An Overview," International Journal of Current Engineering and Technology, vol.4, no.3, June 2014.
[7]. Santosh V. Chapaneri, "Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping," International Journal of Computer Applications, vol.40, no.3, Feb.2012.
[8]. Rashmi C. R., "Review of Algorithms and Applications in Speech Recognition System," International Journal of Computer Science and Information Technologies, vol. 5(4), pp 5258-5262, 2014.
[9]. Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition Using MFCC and DTW," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol.2, Issue 8, Aug 2013.
[10]. J. D. Tardelli, C. M. Walter, "Speech waveform analysis and recognition process based on non-Euclidean error minimization and matrix array processing techniques". IEEE ICASSP, pp. 1237-1240, 1986.
[11]. Takao Suzuki, Yasuo Shoji, "A new speech processing scheme for ATM switching systems". IEEE,Digital Communications Laboratories, Oki Electric Industry Co. Ltd., Japan, pp. 1515-1519, 1989.
[12]. J. S. Lim "Evaluation of a correlated subtraction method for enhancing speech degraded by additive white noise", IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-26, no. 5, pp.471 -472 1978.
[13]. R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," Trans. ASME Series D, J. Basic Engineering, pp. 95-108,1961.
[14]. Gabrea, M.: 'Adaptive Kalman filtering-based speech enhancement algorithm'. IEEE Canadian Conf. on Electrical and Computer Engineering, 2001, vol. 1, pp. 521–526.
[15]. Jeong, S., Hahn, M.: 'Speech quality and recognition rate improvement in car noise environments', Electron. Lett., 2001, 37, (12), pp. 800–802.
[16]. Ma, J., Deng, L.: 'Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model', IEEE Trans. Speech Audio Process., 2003, 11, (6), pp. 590–602.
[17]. Mathe, M., Nandyala, S.P., Kishore Kumar, T.: 'Speech enhancement using Kalman filter for white, random and color noise'. IEEE Int. Conf. on Devices, Circuits and Systems (ICDCS), 2012, pp. 195–198.
[18]. Yeh Huann Goh, Paramesran Raveendran, Yann Ling Goh, "Robust speech recognition system using bidirectional Kalman filter", IET Trans. Pp. 1751-9675, 2015.
[19]. Tony Lacey, "Likelihood interpretation of Kalman filter", tutorial lecture, April 2006.
[20]. Siva Prasad Nandyala, Dr.T.Kishore Kumar,'' International Journal of Computer Applications'' Volume- 12, pp.0975 – 8887, November 2010.
[21]. Paliwal, K.K, Atal, B.S. Efficient vector quantization of LPC Parameters at 24 bits / frame. IEEE Trans. Speech Audio Process, pp.3- 14, 1993.
[22]. W.B. Kleijn and K.K.Paliwal, "An introduction to Speech coding," Speech coding and synthesis, Elsevier science, 1995, pp.1-47.
[23]. J. Makhoul. S. Roucos. and H. Gish, "Vector quantization in speech coding," Proc. IEEE. vol 73, pp. 1551-1588, Nov.1985.
[24]. Sharon Gannot, David Burshtein , Ehud Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms" IEEE Transactions on speech and audio processing, vol.6, pp. 373 – 385, july 1998.
[25]. Zenton Goh , Kah-Chye Tan , Tan, B.T.G., "Kalman filtering speech enhancement method based on a voiced and unvoiced speech model", IEEE Transactions on speech and audio processing, vol.7, pp. 510 – 524, August 1999.
[26]. Ki Yong Lee and Souhwan Jung, "Time–Domain Approach Using Multiple Kalman Filters and EM Algorithm to Speech Enhancement

with Nonstationary Noise", IEEE Transactions on speech and audio processing, vol.8, pp. 282 -291, 2000.
[27]. Lee, Ki Yong, Lee, Ki Yong, "Recognition of noisy speech by a nonstationary AR HMM with gain adaption under unknown noise", IEEE Transactions on speech and audio processing, vol.9, pp. 741 – 746, 2002.
[28]. Chi-Chou Kao, "Design of Echo Cancellation and Noise Elimination for Speech Enhancement", IEEE Transactions on Consumer Electronics, Vol. 49, No. 4, NOVEMBER 2003
[29]. Ma, J.Z., Li Deng, "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state space model", IEEE Transactions on speech and audio processing, vol.11, pp. 590 – 602, January 2004.
[30]. www.developer.android.com, MediaRecorder, State Diagram of Lifecycle of MediaRecorder.
[31]. Yang, C.H.. "A mobile communication aid system for persons with physical disabilities", Mathematical and Computer Modelling, 200802.
[32]. "Computer Networks & Communications (NetCom)", Springer Nature, 2013.
[33]. Neha Sharma, Shipra Sardana. "A real time speech to text conversion system using Matlab", 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.
[34]. "Proceedings of International Conference on Communication and networks", Springer Nature, 2017.
[35]. Eun, Yun-kyu, and Chul-Jin Kim. "A Research and Development of Dynamic Recognition Technique for Enhancing Reliability of Mobile Sensing Service" , Journal of the Korea Academia-Industrial cooperation Society, 2015.
[36]. A.Muthamizh Selvan. "Word Classification Using Neural Network", Communications in Computer and Information Science ,2011